EVALUATION AS PART OF A PROJECT LIFE: THE HYPERMEDIA CAMILLEPROJECT

Thierry CHANIER
Laboratoire d'Informatique de Besançon
Université de Franche-Comté , France
C. E.:thierry.chanier@univ-fcomte.fr

Ref : Journal of the Association for Learning Technology (ALT-J), vol 4, 3, 1996. pp 54-68.

Abstract: In the Computer-Aided Instruction (CAI) literature, the issue of integrating evaluation in the project life has often been recommended but much less reported, at least for large-scaled hypermedia environment. Indeed CAI developers face a difficult problem because they have to make apparently contradictory viewpoints of the audience of the evaluation (teachers, administrators, research community, sponsors) coexist. This paper recalls the various functions and forms of evaluations adopted by the different kinds of audiences. It then reports evaluations, formative as well as summative, set up by the European CAMILLE project teams in four countries in the course of their courseware development. It stresses the advantages, despite drawbacks and pitfalls, for the CAI developers to systematically undertake evaluations. Lastly it points out general outcomes on learning issues, which are of primary interests to teachers, trainers and educational advisors, about topics such as multimedia, learner's variability, autonomy, and effectiveness of learning with respect to language skills.

1. INTRODUCTION

This paper reports on evaluations undertaken in the countries which participated in the CAMILLE project[1]. The principal aim of this European project has been the development and delivery of hypermedia courseware in Dutch, Spanish and French. The courseware encompasses the training of general linguistic competencies for beginners (of Dutch and Spanish) as well as the competencies related to the use of language for specific purposes (French). The target audience include students in science or business, and technicians or engineers from SMEs. This report may be of interest to two kinds of audience. Firstly, each one of our packages relies heavily on the hypertextual and every sort of multimedia facility, which is provided by standard computing platforms. It offers learners a large-scale learning environment, which can assist them in various autonomous situations. Consequently preliminary outcomes of the way CAMILLE has been practically used by the learners, and of measures of its effectiveness concerns teachers, trainers and educational advisors. Secondly, the various experiments have been managed by teams which were working upon the software developments. In the Computer-Aided Instruction literature, the issue of integrating evaluation in the project life, i.e. either in the course of the development or at the end of it, has often been recommended, but much less reported, at least for this type of environment. We discuss here the constraints, advantages and drawbacks of having such commitments.

For the sake of the presentation of the different kinds of experiments achieved in CAMILLE and for the reliability of the interpretations of our partial results a preamble on evaluation is necessary. Evaluation is so widely used by the various audiences connected to Computer Assisted Language Learning (CALL), which have distinct perspectives in mind, that delivering results may lead to misinterpretations. At one end of the spectrum there is an increasing pressure upon researchers and developers teams to adopt more methodological and scientific procedures, and, at the other end, educational advisors and executives constantly require concrete and positive results before extending their support to CALL. CAMILLE is one project, among an increasing number of others, which had to try to make these apparently contradictory

viewpoints coexist. Section 2 describes the various aspects of evaluation in language learning and in CAI.

Section 3 recalls the initial requirements and achievements of the CAMILLE project, and introduces the common features of the different experiments. Section 4 details evaluations made in two countries. Section 5 reports the main general outcomes and sums up our experience of the management of evaluations during a project life.

2. PREAMBLE ON EVALUATION

In order to delimit the framework we adopted for this research, we discuss in this section the principal functions of evaluation, the initial questions in the design process, the overlapping forms of evaluation and the evaluation procedure.

2.1. Functions of evaluation

For almost thirty years a distinction has been frequently made between two principal functions of evaluation: the **formative evaluation** and the **summative evaluation**. This distinction exists in language teaching (Lussier, 1992) as well as in CAI (Knussen & al, 1991 ; Demaizière& Dubuisson, 1992 ; Mark & Greer, 1993), but they are differently interpreted.

In language teaching, the formative evaluation consists in regularly diagnosing the learner's state of knowledge, abilities, and attitudes. It is made for the learner in order to let him/her know his/her current position with respects tithe ultimate goals, and for the teacher to bring information that could help to adjust and adapt his/her teaching, before the end of the course. In CAI, formative evaluation also occurs much before the end of the implementation phase. It is intended to help the designers appreciate the progress towards achieving the goals of an educational innovation. It is set up by designers and involves few learners who are under strenuous observations in order to appreciate whether they use the software as it is intended to be used. Such aspects as interface, person-machine interaction, learner's strategies, hardware configuration and computing architecture are observed with rather informal methods. Outcomes of this process brings detailed, as well as general information, which may provoke surface changes (correction of bugs) or more profound changes in the design and the development. It also provides insights into the way it is going to be integrated into a real-life learning situation.

In language teaching, the summative evaluation comes out on the conclusion of course, or a programme, in order to measure the level of proficiency acquired by a learner with respects to normative goals, explicitly fixed by the learning institution. It is a global measure which compares the performances of learners. It is intended to certify the learner in order to give him/her credits, to recommend an orientation, or to check the effectiveness of the course or programme. In CAI, summative evaluation is concerned with the evaluation of completed systems. Its purpose is to measure the effectiveness of an innovation in terms of the stated aims. It is intended for trainers, centers and designers to assess the suitability, for certain tasks and users, of the software, or to compare it with other products already in use. In both cases summative evaluation has to be undertaken in real-learning settings and to involve a larger number of subjects than the previous formative one.

Here we will adopt the CAI standpoint, and not the language teaching one, when describing evaluations since computing environments developed in the CAMILLE project have been designed as freestanding supports (without any, or with limited, teacher involvement), and the

question of performing evaluations when developing a software package is the central topic of this paper.

Beyond the discrepancies between CAI and language teaching evaluations, there is a common feature which distinguished them from assessment. Evaluation is not a judgement but a decision-making process. Since outcomes may be interpreted by various audiences (e.g. designers, teachers, institutions) in order to make lasting changes, the framework for setting up an evaluation and its procedure will be examined hereafter.

2.2. Initial questions in the design of an evaluation

Evaluating a language program or a CALL software is a complex process. Here are some key questions that should be answered before starting any evaluation (from Nunan, 1992: chap 9):

- Objectives: What is the purpose and who is the audience of the evaluation (for whom is it made)?
- Methodology: What principles of procedure should guide the evaluation? What tools, techniques, and instruments are appropriate?
- Material constraints: Who should carry out the evaluation? When should it be carried out? What is the time frame and budget?
- Release: How should the evaluation be reported?

It may seem obvious that it is extremely important to clarify, from the beginning, the goals of the evaluation. However it is not a straightforward task. Let us consider an innovation. Relationships are not clear at all between the original working hypotheses of designers, the actual achievement, and the selection of precise experimental variables: a shift may have appeared between the starting and end points; an innovation may have unexpected effects (it may not raise the level of proficiency, but the learner's motivation); comparisons with other existing learning environments may be problematic because of their wide discrepancies. For example, determining a scale for measuring effectiveness with respect to communicative goals and specific purposes was an expected outcome, per itself, in the CAMILLE project. Fixing an objective to the evaluation is again not always easy when the audience is diversified: designers, teachers, administrators, funding bodies often have different perceptions.

If learning objectives need to be elicited, they also need to be associated to precise **forms of evaluations** which are themselves associated to different methodological approaches. Here we extract overlapping forms of evaluations, from one out of many possible presentations, (Knussen & al, 1991)):

- Experimental: a limited number and clearly defined variables are scientifically measured, usually based on statistical inferences. The laboratory is the traditional setting for evaluations which generally have affirmative function. If such a form is considered as more scientific, its relevance to real learning settings is problematic.
- Research and developmental: the purpose is to apply quasi-experimental methodologies, including pre and post-tests, in situations closer to real learning settings. This form, which many evaluations of CAL systems try to adopt, also requires clear statements of measurable objectives. They may be easier to guarantee in scientific or industrial environments than in educational ones. They more often concern summative than formative functions.

- Illuminative: isolating variables and associated parameters, quantifying measures is hard to achieve in real learning settings, especially if estimation of the impacts of social factors and the participant's views on the meaning of educational innovations are at stake. Consequently methods, essentially qualitative, usually based on observations and interviews, are applied to "illuminate" important factors rather than to test hypotheses. Pitfalls of this form range from the risk of the observers' obtrusiveness to findings which cannot generalize other settings.
- Teacher as researcher: since teachers play a prominent role in the integration of CALL systems into the curriculum, it sounds natural to let them take in charge the evaluations. Biases (e.g. subjectivity, role-conflict, work overload) introduced by this form suggest that it should only be used in addition to other approaches.
- Case studies: understanding the effects of situational and personal factors in the use of innovative software is generally based on the detailed study of a restricted number of learners. However a generalization of the findings to other situations may be difficult.

In CAMILLE, two forms of evaluation have mainly been used: the research and developmental one, and the illuminative one.

Once the form of the evaluation has been determined, the material constraints need to be appreciated before performing the evaluation procedure. The first step of the procedure, sketched in figure 1, consists in designing the whole evaluation. The initial task of the second step is the construction of the instruments: materials for the tests, questionnaire and forms, and extra materials for the control group, if necessary. Then data collection and analysis is followed, according to the methodological approach chosen. The third step, drafting the report, learning and deciding from it, may not be the last one. In formative evaluation, immediate decisions may be taken, and followed by changes which then will be measured a second time.

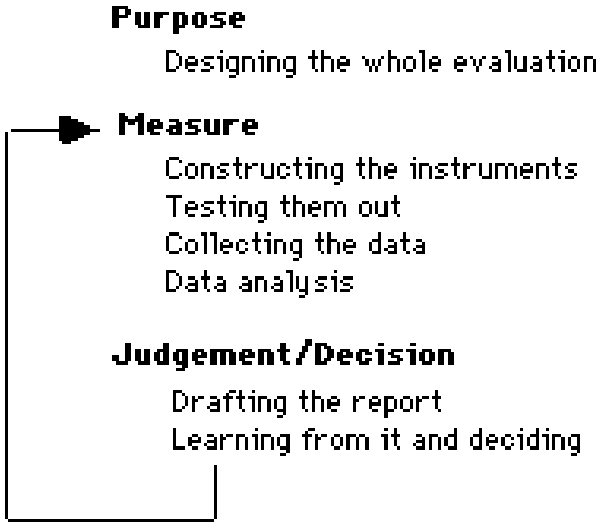

Figure 1 The evaluation procedure

The variety of tasks, and their coordination put forward the availability of realistic material constraints (Who will the evaluators be? What are the time frame and the budget?) for the successful completion of the project. This may explain why CALL developers may be reluctant to perform an evaluation procedure as part of their project life.

The last issue, raised in the initial questions, refers to the release: how is the evaluation to be reported? On one hand, evaluation is often described as a public act which should be open to

inspection, on the other hand, dissatisfactions in the findings, and/or disagreements between participants may impede the final report to see the light of day. On the contrary, interesting findings may be overgeneralized if the final conclusions are not clearly delimited.

3. GENERAL ASPECTS OF CAMILLE EVALUATION

The European LINGUA CAMILLE project started in 1993 and will finish in 1996.Descriptions of the project and of its theoretical standpoints can be found in (Ingraham, Chanier & Emery, 1994; Chanier, 1996; Pothier, 1995). In this section we only recall its initial requirements and its main achievements. From there, we examine the various purposes and audiences of the evaluations, and the common features shared by the different experiments. Details of evaluations are presented in the next section.

3.1. Initial requirements of the whole project

The CAMILLE project is aimed at conducting a large-scale experiment touching upon issues arising from both pedagogical and software engineering viewpoints. From a pedagogical viewpoint, hypermedia technologies are often presented as an opportunity to enhance language learning. Although these factors are often assumed to play an important role in the acquisition process, as yet there have been no large-scale experiments based on their use. CAMILLE was thus seen as an opportunity to undertake such an examination in a multi-cultural environment. The objective was the construction of an environment that would provide the learner with all the tools and information, short of a live teacher, that they might need to undertake a specific level of course in the target language. One consequence was the integration of books/resources (on lexicon, culture, function, grammar) with the textbook (the course proper) on the same desktop. Another consequence was the mode of its use and of its integration into a whole curriculum: CAMILLE was designed to be used by well-motivated adults, who may or may not be engaged in formal education or training and who may or may not have access to a tutor. Thus the emphasis was put on autonomy.

From a software engineering viewpoint, hypermedia programming tools are often recommended as an opportunity to speed up courseware development, and therefore make CAL a realistic complement for training learners in and out of the academic world. But production of courseware in hypermedia also dramatically increases the number of qualifications required and up to now our experience inner-using modules of software, or shared knowledge for large-scale software is still very limited. The CAMILLE project was supposed to help to gain a clearer understanding of trans-national courseware development. As a starting point, it was decided to use a common template for developing, a template which consisted of a software and hardware platform, created by our English partners in 1991/92. The effectiveness of the software engineering viewpoint was enforced by the decision to launch a commercial release at the end of the project, i.e. beginning of 1996.

3.2. Main achievements

A few months before the conclusion of the project, the main courses finalized or near completion are:

- "Español Interactivo", "Interactif Nederlands", and "FranceInteractive". Three packages respectively designed for the training of general linguistic competencies for beginners in Spanish, Dutch, and French, and developed in Spain, Netherlands, and England.

- "Travailler en France". This package has been designed for the training of competencies related to the use of Language for Specific Purposes (LSP) for intermediate-advanced level of French, and developed in France.

Every package includes two CDs which run on a standard and basic hypermedia PC platform, namely the international standard MPC2. It has the minimal equipment to play full motion video and offers good quality for recording and playing sounds. Every disk gathers approximately 30 minutes of original video, plus other oral, graphical and textual data, on top of which are built resources, and several dozens of activities, which offer more than 20 hours of study to the learner.

Whilst, at present, debugging and some coding processes are still under way, CAMILLE partners are fixing the legal aspects in order to start the commercial release of the most advanced courses in 1996. Figure 2 sketches the various overlapping tasks that every partner has been pursuing, and displays the place of evaluations within the project life.

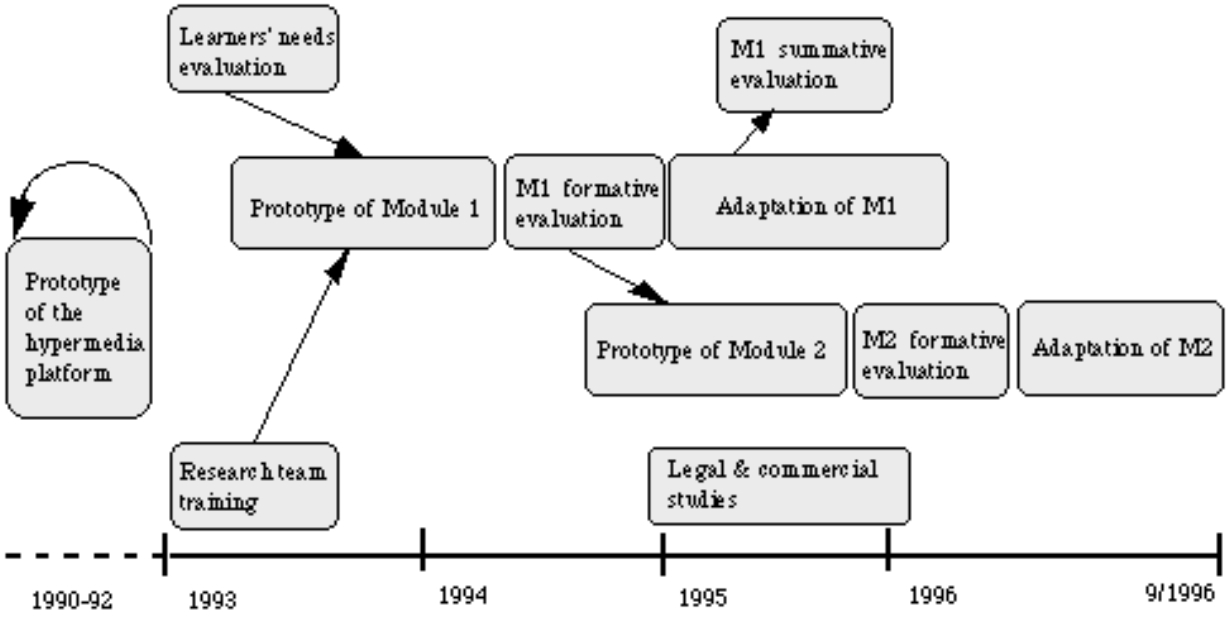

Figure 2 The actual software development process of one course in CAMILLE: the case of "Travailler en France".

3.3. Common features of evaluations

Following the general framework discussed in section 2, we review here the common features of evaluations undertaken by all the CAMILLE partners. Three kinds of audiences with their respective purposes can be distinguished. The first kind of **audience** encompasses the European Union and publishers, as external (to the CALL community) actors which intervene in the project life. The former partially funded CAMILLE (actually for less than a quarter of the total budget), added its own requirements and annually examined our achievements before deciding any extension of funding. The latter, i.e. publishers, have recently undergone internal restructuring in order to be prepared to release multimedia software. Most of them limit the major risks linked to these innovations by expecting developments to be supported by small and recent private ventures. Furthermore they are not accustomed to deal with academic institutions. For them all, evaluations were intended to assert our reliability, by proving that learners could turn their hand to our courseware in real settings, by convincing them that

academics could challenge private companies, and be more transparent when performing evaluations as public acts open to inspection.

The second kind of audience is the CALL community, which includes teachers and researchers. The pedagogical perspectives recalled in the previous sub-section needed to be estimated. Purposes here are twofold: firstly, measuring what kinds of language skills multimedia technologies can help practice, what sorts of learning strategies are performed in hypermedia environments, and how effective is autonomous learning in various settings ; secondly, scaling effectiveness with respects to communicative goals and specific purposes. The latter point refers to the problem of finding criteria by which educational objectives can be measured: how can we assess the learner's ability to master knowledge and skills, mobilized around the specific purposes of every courseware, to transfer them and create new pieces of discourse (cf. Landsheeres' trilogy (de Landsheere, 1984)).

Developers constitute the third kind of audience. The purpose here was to appreciate to what extent is formative evaluation necessary for adapting and debugging the learning environment, and summative evaluation to clarify the software goals, i.e. exactly identify what can be measured, and what are the constraints and overload brought upon the whole project.

The different CAMILLE research teams set up evaluations, located either in their own institutions or in surrounding ones. This happened over 14 months (autumn 1994-1995), during implementation or at the end of large parts of it. No extra budget, nor extra human resources were available. Evaluations are being reported in three different ways: the final report to the European Union, conferences such as EUROCALL (Emery & al, 1995) and scientific papers.

4. DETAILS OF EVALUATIONS

In the CAMILLE project, objectives and methodologies varied from one research team to another. As an illustration, in this section we detail evaluations undertaken on "Interactif Nederlands" and "Travailler en France". Results drawn upon FranceInteractive and "Español Interactivo" are included in the general outcomes presented in the next section.

4.1. Evaluation at HEBO school (Holland)

The Dutch CAMILLE team prepared a double function evaluation, evaluative and formative. The formative side of the evaluation was designed as a two round experiment, thus illustrating the cycle drawn in figure 1. As soon as data was analyzed, changes were made and new experiments were based on the modified software. The aim of the summative evaluation was to compare the software with local classroom learning. This second side directly interested the HEBO managers and the local teachers. The school supervised more than a thousand university Dutch or foreign students who needed intensive training in several languages for professional purposes (legal or business). It offered a strong integration of CALL into the curriculum: nearly 50% of the students' work time in language learning was organized around free access to computers. Heads of the school consider the familiarization with the Dutch language and society by foreign students as an important factor of integration into a country where they are spending several years. Of course, Dutch is not a "survival" language (learners can easily talk English and be understood by anyone in everyday life), but attendance in Dutch classrooms is strongly encouraged, however not mandatory, and learners' credits can easily be transferred. It was thus decided that learners who will only learn Dutch through Interactif Nederland would pass the same usual oral exam as the other learners of Dutch who attend the classroom.

The experiment took place at HEBO School, in the multimedia, free-access room. Evaluators used a network version of the software on computers which had a hardware configuration as we recommended. Sixty local students were involved, on a voluntary basis. They were true beginners in Dutch, experienced language learners (Dutch often being their third language), had a low motivation for learning Dutch, and a basic experience with computers. Learning tasks were organized around half of the software, which represented 30 hours of work, distributed over 10 weeks, with free access conditions. Learners had to fill in questionnaires and were interviewed at the end of each session. Evaluators also made nonsystematic observations.

Data from fourteen students have been analyzed for the formative part of the evaluation. This issue will not be detailed here but what lesson evaluators learned from this experiment will be mixed with the other general outcomes in the next section. The final exam was organized by the usual Dutch teachers, not by evaluators. Marks and teachers' comments on the CALL group showed that results were neither better nor worse than usual. Since the timing and the assessment procedure were the same as the live course, the software has proved to be efficient with these types of situation and learners.

4.2. Evaluations in CAVILAM (France)

Formative and summative evaluations of module 1 of "Travailler en France" have been organized at two different stages of the project (see figure 2): the formative evaluation in October 1994, at the very end of the development of the prototype of module 1; the summative one in June 1995, after changes and debugging had been achieved on module 1 and while module 2 was underdevelopment. We first present the common features of both evaluation, before detailing them.

Experiments were located, outside the university, at the CAVILAM, in Vichy, a semi-private institution specialized in intensive language training for adults. The language laboratory of the institution is only equipped with a few multi-media computers of limited hardware configurations. We used the CD version of the software. Local students, who were following full-time language training periods of 1 to 6 months in length, participated in the evaluation. They were between 21 to 47 years of age, with an average age of 25, coming from various continents and cultures. All were intermediate (200 hours) or advanced (400 hours) learners of French, with French often being the third language. They had good professional motivation for learning, either because they already had a job, or were seeking work, where the mastering of specific skills in French was important, or because they wanted to attend French universities. They had a mixed experience with computers, some being almost illiterate, coming from countries where computers are not part of the natural environment.

Both evaluations have been undertaken with module 1 "A la recherche d'un travail", where the specific purpose is to learn how to apply for a job in France. This makes a noticeable difference from other CAMILLE courses which are for general purposes (Chanier, 1996). This CALL course is built around one main task, making a job application. Knowledge bases and activities allow the learner to fulfil the task and immerse the learner in a socio-cultural context which determine the architecture of the software. The story line of the module presents two characters who are very different in nature and representative of situations subsequently encountered: where to find appropriate information and acquire experience in the employment market, how to write a letter of application and a CV in the French way, how to fix an appointment on the telephone, how to handle an interview. Linguistic knowledge and activities have been designed from the task context, but do not have top priority. The learning tasks require a total amount of

20 to 25 hours work over 3 weeks. The learners use the software during the time usually allocated for practical working their training and had extra free-access possibilities.

The formative evaluation

One purpose of the formative evaluation was the usual measure of the progress of the innovation. The second one focused on the adequation between, on the one hand, the kinds of activities and resources available and, on the other hand, the learners' strategies and interests. The sample population was limited to five volunteers, because we wanted one of our observers to always be present. They could work alone or in a group. The observer, who acted in anon obtrusive way, either video or audio-recorded all the sessions, took detailed notes on their moves, selections, and timing. Learners filled in pre and post-questionnaires and had a form to fill in at the end of every session, followed by a short interview.

Of course, thanks to this procedure, we have been able to collect detailed information of the learners' behaviors and reactions, not only of their (positive) comments, which helped us to make some adaptations. Details will be discussed in the section on general outcomes of the overall CAMILLE evaluation. One point is worth mentioning at this stage, because it relates to the LSP aspect of our software: even when learners were not directly, personally concerned by the task situation they all (even subjects of the summative evaluation) expressed that this experience has represented for them an important discovery of the socio-cultural aspects of the L2 country, and of the everyday native language related to it. Apprehending variations in the target language and links between language and complex situations encountered daily by natives is an efficient way of raising language awareness, which is an important aspect of second language learning.

The summative evaluation

The purposes of the summative evaluation were threefold:

- assessment of suitability for the first LSP courseware with respect to the local learners,
- comparison with autonomous (audio + paper) learning,
- Measurement of the impact of hypermedia CALL on vocabulary learning.

For this second experiment the audience was not limited to the project team. The CAVILAM staff was also interested in the outcomes and took over the supervision of the learning tasks. They acted as counsellors. The project team only handled the various tests.

Subjects were divided into two groups on a voluntary basis: group 1 (G1) , the paper and audio based, comprised of 6 people ; group 2 (G2), the CALL one, 7.For group 1 we extracted large parts of textual data contained in the software activities and resources, and all the sounds of the dialogues. They then had a document and audio-cassettes to work with. They also had access to paper-based dictionaries available in the language laboratory.

We prepared pre and post-questionnaires, the post-questionnaire contents being different for G1 and G2. We also translated into French and administrated the SILL pre-test. The Strategy Inventory for Language Learning test (Ehrman &Oxford, 1990) allows learners to clearly indicate what sort of strategies they usually apply when learning a language generally. Results show to what extent they use (and are away of using) appropriate strategies for remembering more effectively, using mental processes, compensating for missing knowledge, organizing and

evaluating one's learning, managing emotions and learning with others. Subjects also passed a pre and a post-test on vocabulary (pre and post-tests were identical) and a post-test to assess communicative competence on the same domain. For the latter one, called the main post-test, we created original aural and textual materials. Subjects had to write their answers and essays. The main post-test had three parts: an aural comprehension of an interview which included subjective appreciation of the applicant's situation; a comprehension and a written production of part of the exchanges in a dialogue on the telephone; and the writing of a letter of application for a post-profile described in an advertisement. This test was not ready when the experiment started, so we could only use it a post-test.

Figure 3 displays average results per group. The two groups appeared not to be fully balanced. Analysis of subjects' answers in G2, the computer-based group, showed that they used more varied strategies and were more self-conscious of the way they usually learned. They proved to have better lexical knowledge than G1 in the pre-test. Both groups progressed in this domain, G1 slightly more than G2. This may not be very surprising since the lexical test was difficult (the emphasis was put upon the relationship between words and phrase, and collocations; semantic relationships, grammatical structures and relational constraints of lexical phrases were required to be understood).Within this context, progression in subjects with lower level knowledge is easier. As regards the main post-test, diagrams show there was not much difference between G1 and G2. This result is not straightforward to explain in a confident manner since samples were limited in both groups. However we noticed that G1 behave as if they were competing against G2. The learners inG1 did however have to find by themselves extra resources which were given for free in the software: for example, we observed that G1 learners frequently used dictionaries. G1 strongly protested against their learning materials which they found boring, while G2 found much interest in the software. Learning may have been a harder process for G1, but both groups satisfactorily learned and passed their exams (vocabulary and main post-tests), which was what we were expecting.

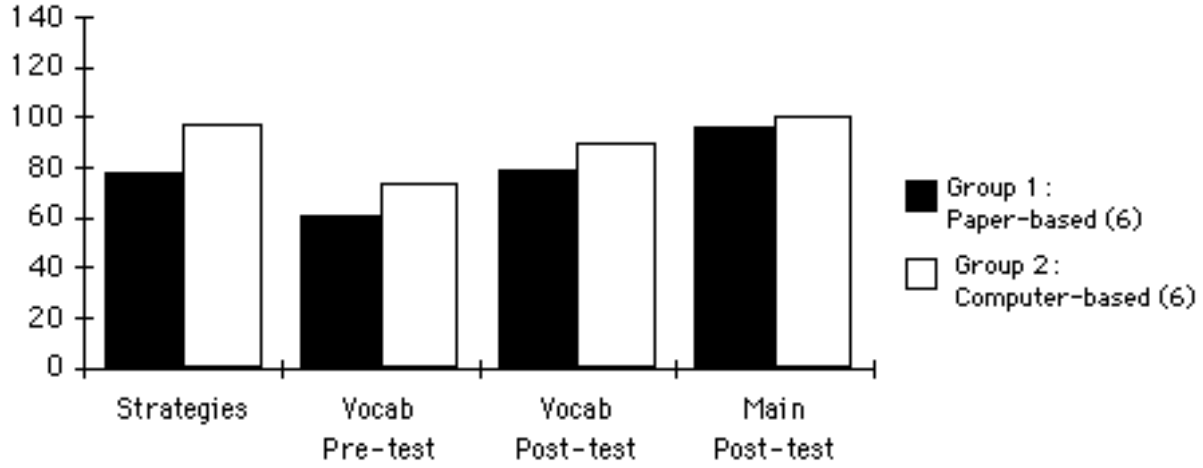

Figure 3 Average results per group
in the summative evaluation of module 1 of "Travailler en France"

5. CONCLUSIONS

Our conclusions will be structured around two topics. Firstly, after having avoided repeating similar issues in the different evaluations, we now summarize the common outcomes. Secondly, we come back to the problem raised in the first sections of the paper, that of undertaking evaluations within a CALL project life, and we explain our viewpoints.

5.1. General outcomes

It is now time to discuss results relating to the common evaluation purposes listed in section 3.3. We consecutively examine issues relating to multimedia, learners' variability, autonomy, effectiveness of learning with respects to language skills, and we try to open new perspectives.

Multimedia

The activities which the learners rated most highly were the video-based and the audio-based activities, in order of preference. When asked to evaluate the activities upon quality alone, this order of preference is reversed. When learners found quality of sound unsatisfactory they expressed their view strongly although they never complained about the definition or size of the video material. This result supports our original decision to use the basicMPC2 standard, since limitations to the quality of video are less important given the range of functions we assigned to video. In CAMILLE, as in other CALL environments, video is primarily used:

- to put language into context, thus to raise motivation,
- to support the interpretation of the linguistic contents of utterances: in simulation activities, looking at the speaker's face may bring information on the pragmatic contents of the message (happiness, irony, discontent, etc.) ;when pronunciation activities are essential, as in the first lessons of the Dutch course, focusing on the speaker's lips movements facilitate comprehension and production of phonemes.

For such functions, the video supports sound. Which means that, when use of video is suppressed (e.g. in some telephone-based activities we wanted to increase the level of difficulty), the linguistic content is still comprehensible, provided that the quality of the sound is very good.

However, learning a language is not reducible to purely linguistic knowledge. Kinesics and proxemics are also very important (Feldman & Rivé, 1991). In real communication settings, the hearer not only interprets the speaker's message from its linguistic content, but also from his/her gestures, location, etc. In such situations as interviews or negotiations (in business and such like), the issue not only relies on what is said, but also fundamentally on the predisposition of the various parties, predisposition which will be interpreted accordingly to a protocol of behavior and gestures. In foreign language learning these aspects are never neglected in live courses. If we want to do as well in CALL, we need to study other functional uses of video. In one module of the French for business courseware, we have designed three activities on gestures, which can either only support the verbal message or completely replace it. No experiment has been made around this new type of activity because its development was achieved after the evaluation phase.

As regards sound, results of experiments showed that we had underestimated potentialities of simple technologies which allow recording and producing sounds of high quality. In CAMILLE it is possible for the learner to record him/herself in almost every activity. In some of them, self-recording is accessory, but in others (like simulation activities) it is essential. Experiments showed that, even if all learners found it important to have self-recording facilities, there existed a large discrepancy between their claim of using them and the extent to which they actually did. This can be explained by the learner's lack of self-insurance, and by the lack of explicit stress, in the first versions of our software, of this important and preliminary step for the support of oral production skills. We now have switched to simple solutions such as adding signposts and

interactive comments in relevant activities and in the general learner's follow up. The role of the computing learning environment is, in fact, to suggest that every user in an interactive way should have recourse to good strategies, such as trying rerecording themselves several times and make a (subjective) comparison with the model (as we observed some learners actually performing the task).

Learners' variability

In all the questionnaires, learners almost unanimously expressed their preference for interactive activities, compared to more passive ones. But they disagreed upon which ones they considered to be better (with the exception of simulations, which were always highly rated). Learners also often stressed the fact that, if they founded communicative activities attractive, basic linguistic activities, on grammar or vocabulary, should not be forgotten. This permitted, in the case of Interactif Nederlands for example, to adjust the balance between both types of activities by adding new, more linguistically oriented, tasks. The learners' reaction was not a plea for activities of "traditional" nature. Linguistic activities can be designed in new ways. For example, learners found our presentation of vocabulary knowledge in lexical networks in "Travailler en France" of great appeal.

Learners' variability appeared not only in their opinions but also in their ways of working with the courseware. They followed very different routes in the scheduling of their overall work: some undertaking activities strictly in the order suggested by our presentation ; others getting a quick overview of the whole contents and of the various kinds of activities (which were signposted), then starting by the ones they a priori preferred, and achieving the rest afterwards. Subjects also performed every activity in very different ways: some trying to quickly finish it without paying much attention to instructions, without looking at the associated resources (they generally then got stuck and had to restart) ; others self-monitoring their task by first carefully thinking in which order they should proceed, looking at the cues, and available resources ; some being systematic and relying on repetitions of self-recordings, exhausting the various possible alternatives ; some systematically taking notes before actually performing any activity ; some verbalizing their thoughts and reactions, whereas others were almost completely silent. When group work occurred, and when skills were complementary inside the group, effective collaboration took place for sharing responsibilities, one taking over the interaction with the system, while the second controlled the planning, or negotiated the knowledge.

This twofold learners' variability is an important positive outcome. Disagreement on attractiveness of activities showed that everyone found his/her own interest. Variations in the way of using the software happened accordingly to learners' personal features. Whatever our wishes may be by expecting learners to follow the same route, individual variability remains the rule in language learning (Ellis, 1994). One of the advantages of multimedia learning environments is to support these individual variations by offering: different types of activity, practice of different linguistic skills, flexible navigation, and access to resources of various kinds, note-taking.

Autonomy

CAMILLE has been designed for an audience of learners who are typically clients, professionals with clear demands and for whom flexibility and swiftness are essential criteria. Sample populations involved in our evaluations mostly corresponded to this profile. Furthermore they were nearly all experienced learners, either advanced learners in L2, or

beginners in L3.They were aware of their strategies of learning. They used software in an autonomous way, evaluators and teachers, when present, being only observers. From the learners' answers (see figure 4 for detailed figures coming from the CAVILAM, which are representative of the subjects' opinions encountered in other locations) and from our observations, it is possible to underline the following points:

- When we developed our software we pointed out the distinctions between activities and resources, and the needs for resources tightly linked to activities which should bring essential extra knowledge in order to offer self-contents courses (Chanier, 1996). The fact that learners largely used these implies that a large amount of time should be allocated to their development in such kinds of hypermedia environments.
- Software can be self-containing, but still learners are looking for discussion and feedback with experts. It is still an open question to decide whether experts should be teachers, who would act as guides or counsellors, or natives
- Self-access has been, as far as possible, the rule. Learners have made it a basic requirement. Insufficient provision of equipment and of large and flexible time access in institutions may jeopardize the whole learning procedure.
- Autonomous learning situations have only been explored in training institutions. Learners said they were willing to work alone, and to work at home. We do not know whether they would be prepared to pay for this purpose and we have to investigate what the outcomes could then be. Experimenting access at work is still another perspective.

Would you prefer to use the software at	home: 8	work: 4	training institution : 4
Would you prefer to be supervised by an expert?	yes:6	no: 5	
Would you prefer to work	in group: 2	alone:10	
Would prefer to use it	Before starting your training: 0	In parallel with your training: 8	After your training: 3
Are you interested in having such kinds of software available outside any language training institution	yes : 11 (some preferring other topics)		

Figure 4: Learners' answers, after the experiments, on the various ways of using"Travailler en France". 12 subjects were concerned. Multiple answers werepossible. Some subjects did not respond to every question.

Not surprisingly, the types of learners we had very positively reactedto this learning situation. They appeared to master three domains which areessential for being able to manage one's own learning (Holec, 1990):methodological aspects, linguistic aspects and cultural backgrounds inlearning. We have collected no data for generalizing these outcomes to other types of learners. The experiment undertaken in Teesside with true L2 language beginners, lacking self-assurance and motivations recommended here, was not conclusive. Blin (1995) also remarked that an insufficient level of confidence in using computers for language learning purposes (which never appeared to be a problem with our experienced learners) may represent a major element in the learner's decisions to under-use computers as opposed to other

materials in self-access centers. Still much research is necessary for exploring autonomous learning with various types of learners.

Effectiveness / language skills

It is now time to come back to the question of the effectiveness of such hypermedia software with respects to the four language skills. As pointed out the technology we relied upon is more adequate for practicing aural (listening) and written comprehension than aural and written productions. Learners passed2 summative tests, as described earlier on, in HEBO and in the CAVILAM. The former one was completely based on aural skills, thus included aural productions. The later one encompass aural comprehension, written comprehension and written production. In order to correctly appreciate this success, it should be recalled that evaluations happened on small samples (though in HEBO they were much more statistically significant than in the CAVILAM), they related to specific types of learners, and in both places quality of results were not much better than more traditional approaches, live courses or audio-cassette methods. This quality is satisfactory because we were not expecting computing learning environments to be much more efficient, but to represent an effective alternative which can be taken into account in autonomous learning situations, an alternative which possesses other advantages discussed in the previous sections. Another open question is whether or not our results can be generalized to all experienced learners.

5.2. Evaluation as part of the development process

We would like to sum up our experience on the management of evaluations during a project life. We will also stress methodological issues. Let us first begin to consider the formative evaluation, and then, the summative.

The main goal of the formative evaluation is to measure the progress of the innovation. It is a necessity for adapting the software, for debugging it, and for collecting essential information on the timing, the duration, etc.; information which can then also help write the user manual delivered with the software. The procedure must involve real learners belonging to the target audience and should be set up long before the end of the development. The elapsed time between the final release of the software and the evaluation phase often is as long as the duration of the development of the first version of the software which served in the evaluation. In general, a reduced protocol is sufficient. However if research questions are at stake, an extended protocol is necessary for setting up case studies. The whole evaluation procedure becomes then much more complex.

The purpose of the summative evaluation is to measure the effectiveness of an innovation in terms of the stated aims. We have pointed out several caveats: the summative evaluation is time consuming; it requires adequate means for achieving it; many partners are often involved; its results or its abandonment may be used against the project. Since it represents a real risk for the whole project, the first question which should be answered before making any decision is: who is the audience? Who really wants to know the outcomes?

Nevertheless, organizations of summative evaluations by project teams should happen more frequently. They are important for the research and pedagogical communities for deontological reasons:

- They help to clarify the functional differences between various sorts of software reports and the evaluation reports. For most of the software, the only accessible reports are commercial reports, written by publishers, or software reviews, written by external teachers or researchers. These reports may bring useful information, but they have the disadvantages of often being labelled as "evaluations". Such confusion with reports based on experimentsinvolving real learners and following a methodological procedure should beavoided.
- They minimize overgeneralizations, either pro or cons. An evaluation has specific aims. Results can only be interpreted with respect to the restricted parameters which have been tested. Unfortunately, papers are still published which either present evaluations as being aimed at definitely stating the superiority of CALL over other learning methods, without defining parameters such as types of learners, types of skills, levels of proficiency, nor actual learning situations, or papers which, when they make explicit their restricted purposes, do not incorporate any detailed information. It is then impossible for readers to correctly appreciate their results, and to undertake other evaluations in order to verify them.
- They reinforce the idea that in an evaluation, not only the software may be tested, but also the learning situations. A limited software can be very useful, and on the contrary, a wonderful language package can be mis-used, depending on its integration into the curriculum, its access conditions, hardware configurations, etc.
- They may offer instruments for measuring various aspects of the so-called communicative competence. Such references support the dialogue between designers and the Second Language Acquisition community.

Eventually, a summative evaluation is also of direct interest for the project team itself. It represents an efficient way of clarifying the final aims of the software, to estimate the inevitable shift between the initial hypotheses and the reality of the achievements. The creation of instruments of measuring makes it possible to elicit how and on what grounds designers want their innovation to be estimated. Since evaluation is a cumulative process, it forms a starting point from which other researchers are able to set up new experiments in order to extend the initial measures. Tests can also be adequately joined to the software delivery in order to let the learner evaluate him/herself at the end of his/her training period.

ACKNOWLEDGEMENTS

I would like to thank all the teachers and researchers who participated in the CAMILLE evaluations, more particularly: Ana Gimeno Sanz, in Universidad Politecnica de Valencia, Jan Brouwer in De Haagse Hogeschool, Janina Emery,Chris Emery and Bruce Ingraham in Teesside University, the CAVILAM staff inVichy, Maguy Pothier, Paul Lotin and Jérôme Ollier inUniversité Clermont 2.

REFERENCES

Blin F. (1995): "Integrating CALL in the negotiated learner-centred curriculum: a case study". *EUROCALL'95 conference*, Valencia, Spain, September. 19 p.

Chanier T. (1996) : "Learning a Second Language for Specific Purposes within a Hypermedia Framework". *Computer-Assisted Language Learning (CALL)*, vol9, 1, pp 3-43.

de Landsheere V. & G. (1984): *Définir les objectifs de l'éducation*. PUF: Paris.

Demaizière F.& Dubuisson C. (1992): *De l'EAO aux NTF, utiliser l'ordinateur pour la formation.* Ophrys: Paris.

Ellis R. (1994): *The study of Second Language Acquisition*. Oxford University Press.

Emery C., Ingraham B., Chanier T. & Brouwer J. (1995): "Creating Interactive multimedia CALLware: the CAMILLE experience". *EUROCALL'95conference*, Valencia, Spain, September.

Ehrman M., Oxford R. (1990): "Adult Language Learning Styles and Strategies in an Intensive Training Setting". *The Modern Language Journal*, 74, iii.pp 311-327.

Feldman R.S. & Rimé B. (1991): *Fudamental of NonverbalBehavior*. Cambridge University Press et Editions de la Maison des Sciences de l'Homme : Paris.

Holec H. (1991): "Autonomie et apprentissage auto-dirigé, quelques sujets de reflexion". *Les auto-apprentissages, actes des 6èmesrencontres de l'ASDIFLE. Les cahiers de l'ASDIFLE* : Paris , 2. pp 23-33.

Ingraham B., Chanier T., Emery C. (1994): "CAMILLE: A European project to develop language training for different purposes, in various languages on a common hypermedia framework". *Computers and Education* , vol. 23, 1/2, pp107-115.

Knussen C., Tanner G.R. & Kibby M.R. (1991): "An approach to the evaluation of hypermedia". *Computers and Education*, vol. 17, 1. pp 13-24.

Lussier D. (1992): *Évaluer les apprentissages dans une approche communicative*. Hachette: Paris.

Mark M.A. & Greer J.E. (1993): "Evaluation methodologies for intelligent tutoring systems". *Journal of Artificial Intelligence in Education*, vol. 4, 2/3. pp 129-153.

Nunan D. (1992): *Research methods in language learning*. Cambridge University Press.

O'Malley J.M. & Chamot A.U. (1990): *Learning Strategies in Second Language Acquisition*. Cambridge University Press.

Pothier M. (1995): "Travailler en France: un environnement informatique hypermedia pour l'auto-apprentissage sur objectifs spécifiques". *Revue de Phonétique Appliquée*. 14 p. To appear.

Notes

[1] CAMILLE (which stands for Computer-Aided Multimedia Interactive Language Learning Environment) has partly been financed by the LINGUA Programme. Members of the CAMILLE Consortium are The University of Teesside, Middlesbrough, England ; Université Blaise Pascal and Université d'Auvergne, Clermont-Ferrand, France ; De Haages Hogeschool, The Hague, Netherlands ; and Universidad Politécnica, Valencia, Spain.